Combining Goal Models, Expert Elicitation, and Probabilistic Simulation for Qualification of New Technology

Mehrdad Sabetzadeh[1] Davide Falessi[1,2] Lionel Briand[1] Stefano Di Alesio[1]
Dag McGeorge[3] Vidar Åhjem[3] Jonas Borg[3]
[1]*Certus Software V&V Center, Simula Research Laboratory, Norway*
[2]*University of Rome (Tor Vergata), Italy* [3]*Det Norske Veritas, Norway*
Email: {*mehrdad,falessi,briand,stefanod*}*@simula.no* {*dag.mcgeorge,vidar.ahjem,jonas.borg*}*@dnv.com*

***Abstract*—New technologies typically involve innovative aspects that are not addressed by the existing normative standards and hence are not assessable through common certification procedures. To ensure that new technologies can be implemented in a safe and reliable manner, a specific kind of assessment is performed, which in many industries, e.g., the energy sector, is known as Technology Qualification (TQ). TQ aims at demonstrating with an acceptable level of confidence that a new technology will function within specified limits. Expert opinion plays an important role in TQ, both to identify the safety and reliability evidence that needs to be developed, and to interpret the evidence provided. Hence, it is crucial to apply a systematic process for eliciting expert opinions, and to use the opinions for measuring the satisfaction of a technology's safety and reliability objectives. In this paper, drawing on the concept of assurance cases, we propose a goal-based approach for TQ. The approach, which is supported by a software tool, enables analysts to quantitatively reason about the satisfaction of a technology's overall goals and further to identify the aspects that must be improved to increase goal satisfaction. The three main components enabling quantitative assessment are goal models, expert elicitation, and probabilistic simulation. We report on an industrial pilot study where we apply our approach for assessing a new offshore technology.**

Keywords. Technology Qualification, Assurance Cases, Goal Modeling, Expert Elicitation, Monte Carlo Simulation.

I. Introduction

Most systems in critical application areas such as healthcare, avionics, and energy are subject to some form of assessment to ensure that the risks associated with the use of the systems are properly mitigated. The most widely-known type of assessment is certification, conducted by an independent professional or regulatory body, to verify that a system is in compliance with one or more applicable standards. In fast-growing markets, such as the energy sector, assessors are frequently faced with innovative technologies that are not fully addressed by the existing standards and hence are not assessable through common certification procedures. To verify that a new technology will work as intended, a specific type of assessment is performed, which in many industries, e.g., the energy sector, is known as Technology Qualification (TQ). Briefly, TQ is aimed at demonstrating with an acceptable level of confidence that a *new* technology will function within specified limits.

To better illustrate the situations where TQ is applied, let us consider an example from the energy and offshore domain: Steel cables have been used for a long time as the primary apparatus for mooring and installation of floating and underwater structures. Recently, there has been a growing interest in *fiber rope* technologies, both as an alternative to steel cables, and further to enable operations that were previously not possible (e.g., installation in deep water). Existing standards for mooring and installation tend to focus on steel cables. Since steel and fiber have very different physical and mechanical properties, these standards are neither fully applicable to, nor cover the entire set of concerns relevant to fiber ropes. In cases like this, TQ is instrumental (and sometimes mandatory) to ensure that the new technologies can be deployed in a safe, reliable, and environment-friendly manner.

In this paper, building on the notion of goal-based assurance cases [10], we propose a quantitative assessment approach for TQ. Our approach, which is supported by a software tool, includes three main components: *goal models*, *expert elicitation*, and *probabilistic simulation*. We use the the KAOS goal modeling notation [22] to structure and decompose a technology's (safety and reliability) goals. We apply expert elicitation techniques [1], [13] for soliciting expert probabilities based on the collected evidence and for mitigating potential biases. Arguments about dependability generally have a strong reliance on expert judgment [11]. This is also true in TQ. Dependence on expert judgment is particularly strong in early TQ stages where little evidence exists about a new technology. One of the aims of TQ is to identify critical areas where there is significant uncertainty in expert judgments and to define objective fitness criteria to reduce the uncertainty and dependence on subjective opinions. This all makes it important to follow a rigorous expert elicitation process in TQ. Lastly, we use Monte Carlo simulation [18] to measure goal satisfaction and to identify the weak links that must be improved for reducing the uncertainty in the satisfaction of high-level goals.

The contributions of this paper are: (1) Tailoring expert probability elicitation into (KAOS) goal models; (2) augmenting Requirements Engineering goal propagation methods [7] with Monte Carlo-based analyses; and (3) applying the KAOS notation and our augmented propagation solution in the context of TQ. The foundations of our approach are general and can be used for various types of assessment, but the methodological steps of our work are motivated by the workflow of activities in TQ. To encourage industrial adoption of our contributions, we align our work with the guidelines in DNV's Recommended Practice for Technology Qualification [17] and Offshore Service Specification [21].

We report on an industrial pilot study where we applied our approach for assessing an important aspect of the behaviour of fiber ropes in safety-critical offshore systems. Results indicate that our approach offers benefits by making the assessment process more structured and transparent.

The remainder of the paper is structured as follows: In Section II, we give a summary of the TQ process and motivate our work in that context. We describe our approach and its components in Section III. We discuss tool support in Section IV and report on the application of our approach to the fiber rope case in Section V. We compare our approach with related work in Section VI and conclude the paper in Section VII with a summary and suggestions for future work.

II. Background and Motivation

In this section, we provide a brief summary of the activities in TQ (based on DNV RP-A203 [17] and OSS-401 [21]), along with the observations that motivated the development of a goal-based assessment approach for use in this context.

1) ***Specification of Qualification Basis.*** TQ begins with the development of a qualification basis. The basis covers: (1) the technology's main objectives and expectations expressed as functional requirements and parameters, and (2) technical specifications for deployment, operation, and decommissioning of the technology.
2) ***Elaboration of Novel Aspects***. The technology's novel aspects (functions, components, processes) are identified. These aspects are then decomposed to a level of detail at which potential failure mechanisms can be determined, analyzed, and prioritized. This decomposition is performed by qualified experts representing the relevant technical disciplines and fields of experience.
3) ***Planning and Collection of Evidence.*** An evidence collection plan is developed and the plan is executed. Evidence collection activities are targeted at providing quantitative measures, predominantly in probabilistic terms, for the uncertainties and likelihoods of failure. The evidence can, among others, include laboratory tests, theoretical analyses and simulations, procedural changes to avoid potential problems, and tests to reduce uncertainty in analytical models, e.g., erosion models.
4) ***Verification.*** This involves analyzing the qualification basis, the risk studies, and the collected evidence to confirm that the requirements in the qualification basis are met, and that the identified risks are properly mitigated.

We note that while we present the TQ steps in a linear manner, in practice, TQ is an iterative process. This means that before deployment, the technology concept may undergo several rounds of improvement based on the observations and the results at different steps of the TQ process.

The framework we propose in this paper was motivated in large part by the following observations in the current practice:

- **A: Traceability and Rationale.** Verification can be challenging as the assessment body must establish that there is a demonstrable link among (1) the qualification basis, (2) the identified risks, and (3) the collected evidence. A compliance matrix can be used to establish these links, but this approach is limited in that it does not record the reasoning as to why different elements are linked.
- **B: Handling of Expert Judgment.** The process taken to elicit expert judgments, the information elicited from the experts, and the way the information is compiled is not always made explicit. This can have a negative impact on the transparency of the TQ process, thus making it hard for both the assessment body and potential end-users to build the required level of trust in the new technology.
- **C: TQ Costs.** Evidence collection (mainly testing) accounts for the majority of TQ costs. Time and budget overruns may occur if effort is not focused on building or improving the right evidence information. Two issues are frequently raised: (1) Vendors undertake costly tests which turn out to be tangential to the TQ process. (2) Vendors undertake appropriate tests but verification indicates a lack of confidence about whether the TQ requirements are met. In such cases, it is difficult to determine which aspects of the evidence need to be improved because the main factors contributing to the uncertainty about the satisfaction of TQ requirements cannot be easily identified.

In our approach, described in Section III, we use goal models to address **A** by maintaining a logical trace of how TQ requirements are decomposed and linked to the relevant risks and evidence. The decomposition also ensures that evidence collection can be better planned, thus reducing the likelihood of collecting non-useful evidence – see first point of **C** above. To address **B**, an explicit mechanism is incorporated into goal models to elicit, record, and propagate expert probabilities. For the second point in

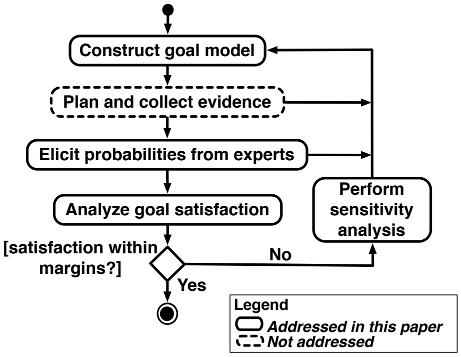

Construct goal model
Plan and collect evidence
Elicit probabilities from experts
Analyze goal satisfaction
Perform sensitivity analysis
[satisfaction within margins?]
No
Yes
Legend
Addressed in this paper
Not addressed

Figure 1. Approach overview.

C, we use a probabilistic technique known as sensitivity analysis [18] to identify the main sources of uncertainty in the satisfaction of a given goal, thus helping focus TQ resources on providing the evidence that reduces the most important uncertainties first.

III. Approach

Figure 1 shows an overview of our approach: we begin with constructing a goal model where we decompose the technology's overall safety and reliability goals, as identified in the technology qualification basis (see Section II), into more concrete subgoals and obstacles[1]. The next step is devising and executing an evidence collection plan. The evidence enables quantifying the following: (1) probability of low-level goals being satisfied, (2) probability of low-level obstacles occurring, and (3) probability of risks arising from incomplete goal decomposition. In our work, we do not aim to provide a specific solution for evidence planning and collection, as this step depends on the technology. Following evidence collection, experts use the evidence to form opinions about the three probability types above.

We then automatically propagate the elicited probabilities to compute probability distributions for the satisfaction of the technology's overall goals. If the level of satisfaction of the overall goals is low, sensitivity analysis will be performed to identify the input quantities with the most significant impact on goal satisfaction. Based on the results of this analysis, we can go back to the previous steps to make improvements, such as including additional provisions in the technology (leading to goal-model updates), using more dependable components in the technology, collecting further evidence, and using additional or more suitable individuals for expert elicitation. The iterative nature of the TQ process is further justification for an explicit argument model, that can be modified, and then used for re-running the analysis.

A. Goal Modeling

We use goal models for decomposing a technology's safety and reliability goals into more specific criteria for which concrete evidence can be collected. Several languages exist for goal modeling; notable examples are: i^* [24], GSN [10], and KAOS [22]. While the main ideas of our approach can be used in conjunction with any of these languages, we choose to ground our work on (a subset of) the KAOS language. This choice is motivated by two main reasons: (1) KAOS formal decomposition semantics is a suitable fit for the type of quantitative reasoning needed in TQ; and (2) KAOS comes with an extended and unified set of modeling guidelines in a book [22]. Such a book is an advantage for training and technology transfer to practitioners.

To illustrate goal modeling in KAOS and further motivate our case study, we use a simplified and sanitized version of the goal model for arguing about fiber rope safety in mooring systems. This simplified goal model is shown in Figure 2. We concentrate on a particular aspect of the fiber rope behaviour that is markedly different from steel cables. In design, the integrity over prolonged time is handled by design curves. The design curve for steel cables describes the safe service life in a relationship between loading range and number of cycles; whereas the design curve for fiber ropes expresses the relationship between static tension and time to rupture. Due to the logarithmic relationship between tension and stress-rupture time for fiber ropes, design considerations have to be made when tension exceeds 75% of the characteristic fiber-rope strength [16]. The focus of our study is to demonstrate that, for a specific type of fiber rope used in a specific environment, safety is not compromised as the result of the rope's time-dependent integrity during a major storm (potentially lasting for several days). The model fragment that is the subject of our analysis is distinguished in Figure 2 with a dashed boundary.

Each goal is "a prescriptive statement of intent that should be satisfied" [22]. Goals in KAOS are depicted in the parallelogram shape (e.g., GL1). The assumptions under which a given goal is to be satisfied are made explicit and captured via an assumption node, presented as a semicircle (e.g., AP1). Goal decomposition is performed using AND and OR operators to show either the case where several subgoals together contribute to the satisfaction of the parent goal, or where alternatives exist for satisfaction. The decomposition can be either full or partial. Full decomposition means that a parent goal has been completely refined and that no more subgoals will be added to the decomposition; whereas partial decomposition means that more subgoals may be added in the future. Partial decomposition is shown using an empty circle and full decomposition is shown using a filled circle. In Figure 2, we use both full and partial decomposition. For example, GL2 is OR-decomposed using full decomposition and GL3 is AND-decomposed using partial decomposition.

[1] Obstacles are events that obstruct goal satisfaction (see Section III-A).

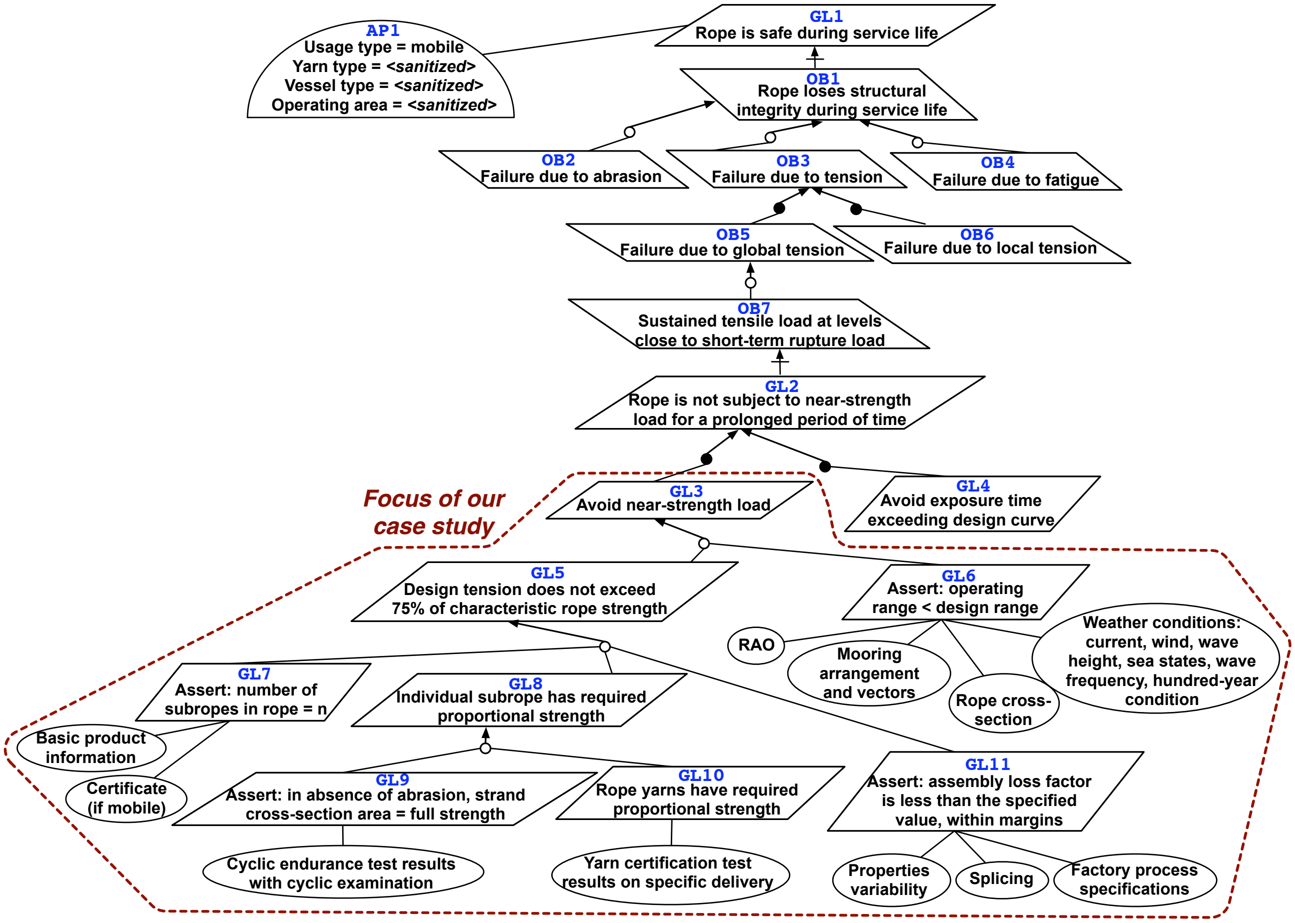

AP1
Usage type = mobile
Yarn type = <sanitized>
Vessel type = <sanitized>
Operating area = <sanitized>
GL1
Rope is safe during service life
OB1
Rope loses structural integrity during service life
OB2
Failure due to abrasion
OB3
Failure due to tension
OB4
Failure due to fatigue
OB5
Failure due to global tension
OB6
Failure due to local tension
OB7
Sustained tensile load at levels close to short-term rupture load
GL2
Rope is not subject to near-strength load for a prolonged period of time
Focus of our case study
GL3
Avoid near-strength load
GL4
Avoid exposure time exceeding design curve
GL5
Design tension does not exceed 75% of characteristic rope strength
GL6
Assert: operating range < design range
RAO
Mooring arrangement and vectors
Rope cross-section
Weather conditions: current, wind, wave height, sea states, wave frequency, hundred-year condition
GL7
Assert: number of subropes in rope = n
Basic product information
Certificate (if mobile)
GL8
Individual subrope has required proportional strength
GL9
Assert: in absence of abrasion, strand cross-section area = full strength
Cyclic endurance test results with cyclic examination
GL10
Rope yarns have required proportional strength
Yarn certification test results on specific delivery
GL11
Assert: assembly loss factor is less than the specified value, within margins
Properties variability
Splicing
Factory process specifications

Figure 2. Simplified goal model for fiber rope mooring.

The obstacles that prevent (obstruct) the satisfaction of goals are depicted as mirrored parallelograms. For example, in Figure 2, the obvious obstacle to the fulfillment of GL1 is that the rope breaks (OB1). This is called the root obstacle. The root obstacle is then decomposed into the factors that can lead to it (in this case, OB2–OB4). Obstacle decomposition is done in exactly the same manner as goal decomposition. Further, just as obstacles can obstruct goal satisfaction, goals can be defined to mitigate the obstacles (e.g., GL2 mitigates OB7).

Goal and obstacle decomposition continues until we reach criteria that are fine-grained enough to be supported by concrete evidence. Evidence items are depicted as ovals and are linked to the relevant leaf goals and obstacles (e.g., see ovals connected to GL6). The standard KAOS language does not provide notational elements for representing assumptions and evidence items. The notation we use for evidence items is borrowed from GSN [10].

Developing a goal model before planning the evidence makes the evidence collection process more targeted and helps avoid activities with limited usefulness, e.g., an expensive full-scale test that despite its impressiveness does not challenge the technology at the operational boundaries. The evidence item(s) linked to each leaf goal (or obstacle) provide experts with information enabling them to estimate the probability of goal satisfaction (or obstacle occurrence). These probabilities are then propagated up the goal model to assess if the overall goals are adequately satisfied. We discuss expert elicitation and goal propagation in Sections III-B and III-C respectively.

B. *Expert Probability Elicitation*

We begin this section by describing the probability types that we need to elicit from experts in the TQ process. Drawing on the existing literature [1], [13], we then provide guidelines for probability elicitation. Our guidelines are intended for the situation where there is uncertainty about the probabilities, and a probability distribution has to be specified for each quantity rather than a point-value.

1) Description of Elicitation Quantities.: There are three types of probabilities that need to be elicited from the experts in our approach:

- *Probability of satisfaction of a leaf goal*: Experts provide the probability of a (leaf) goal to be satisfied. In particular, given a (leaf) goal G and supporting evidence items $E_1, \ldots, E_\ell$, experts need to answer the following: "Based on $E_1, \ldots, E_\ell$, how likely is G to be satisfied?"
- *Probability of occurrence of a leaf obstacle*: Given a (leaf) obstacle O and supporting evidence items $E_1, \ldots, E_\ell$, experts need to answer the following: "Based on $E_1, \ldots, E_\ell$, how likely is O to occur?"
- *Probability of incompleteness risks*: As we discussed in Section III-A, when a goal or obstacle is decomposed, the decomposition may be partial. From a risk assessment perspective, it is reasonable to treat partial *OR* decomposition for *goals* and partial *AND* decomposition for *obstacles* as complete because these kinds of partiality do not impose hidden risks. For example, in the case of partial OR for goals, we are not interested in the probability that a parent goal is satisfied although none of its OR-children have been satisfied. The same applies to partial AND-decomposition of obstacles. However, partial AND-decomposition for goals and partial OR-decomposition for obstacles pose risks. Therefore, given a parent goal G (resp. obstacle O) and subgoals $G_1, \ldots G_n$ (resp. sub-obstacles $O_1, \ldots, O_n$), the experts need to answer the following:
  - *Partial AND for goals*: "How likely is goal G to fail despite all subgoals $G_1, \ldots, G_n$ being satisfied?". We denote the answer by α.
  - *Partial OR for obstacles*: "How likely is obstacle O to occur despite none of sub-obstacles $O_1, \ldots, O_n$ having occurred?" We denote the answer by β.

  We note that, just like for leaf goals and leaf obstacles, one can link partial decompositions to evidence items to support the elicitation of α and β. However, in the industry examples we have seen so far, including our case study, the incompleteness risks were dictated by the rationale for decomposition. While this rationale is critical and needs to be expressed, no specific actions were required during evidence planning and collection. Hence, we do not have evidence items linked to the partial decompositions in the goal model of Figure 2.

For example, to reason about the satisfaction of GL3 in Figure 2, eight quantities need to be elicited: the probabilities of satisfaction for GL6, GL7, GL9, GL10, GL11; and the value of α for the decompositions of GL3, GL5, and GL8. We describe the protocol for conducting the elicitation next.

2) Elicitation of Distributions with Unknown Characteristics.: As we stated earlier, we are interested in eliciting probability distributions from experts so that we can account for the uncertainty. In TQ, the characteristics of the probability distribution to elicit are usually unknown. Among the parametric distributions, the *triangular distribution* seems to be the most suitable option for elicitation, because the parameters of this type of distribution are highly intuitive and easy to respond to for the experts [1].

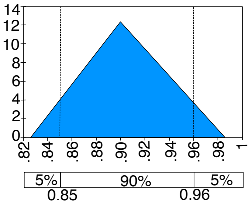


Figure 3. Triangular distribution with extended confidence intervals.

To define a triangular distribution, we need three parameters: the most likely value (m), the maximum (b), and the minimum (a). There is evidence [14] suggesting that the experts should first provide the maximum value. Further, because humans tend to underestimate the maximum and overestimate the minimum, the elicited distributions are often adjusted by enlarging the confidence interval [6]. A straightforward approach to adjust the triangular distribution is proposed in [6] where both end points are extended by equally distributing a certain percentage of the distribution before the minimum and after the maximum. Figure 3 provides an example of triangular distribution with the end points extended by 5% in each direction. The values provided by the expert here are $a = 0.85$, $m = 0.90$, $b = 0.96$. By distributing 5% of the mass around the ends, we get a triangular distribution with $a = 0.82548$, $m = 0.90$, $b = 0.9863$.

3) Elicitation Protocol: Our elicitation protocol is based on the guidelines in [13] and consists of four main steps, which we briefly describe. Due to lack of space, we omit several details, including criteria for expert selection, time and budget considerations, and preparation of elicitation instruments.

- *Step 1 (Recording Experts' Information)* For each expert, we record: name and contact information, field of work, degree, and years of experience.
- *Step 2 (Introduction)* We brief experts about the aim of the elicitation, and prepare them by showing examples and highlighting common mistakes. We further make experts aware of the potential intrusion of bias, see Step 4 below.
- *Step 3 (Soliciting answers)* For each quantity, the following process is used:
  - Read the description of the quantity to be elicited.
  - Have the expert: (1) explain their understanding of the relevant evidence (or rationale), (2) recall the possible operating conditions and circumstances under which the quantity can be assessed, and (3) relate the setting under which the evidence was

collected and the situations arising in practice.
  - Ask: "From your experience and based on the existing evidence, in which operating conditions and circumstances would the probability be very high? What would then be the max. probability?"
  - Ask: "From your experience and based on the existing evidence, in which operating conditions and circumstances would the probability be very low? What would then be the min. probability?"
  - Ask: "Is the most likely value closer to the min. or the max.? What would you deem the most likely value to be?"
  - Define a triangular distribution, based on the elicited max., min., and most likely values, extending the max. and min. values by a pre-specified confidence interval (see Figure 3). We used 5% for extending the interval.
- *Step 4 (Handling bias)* During elicitation, the interviewer should monitor the experts' verbalized thoughts and body language, as well as the group dynamics (if a group setting is used for elicitation) for signs of bias. Table I summarizes the biases most relevant to TQ, along with mitigation strategies. For a more thorough overview of elicitation biases, consult [13].

C. Analyzing Goal Satisfaction

We propagate the values obtained through expert elicitation to compute a satisfaction distribution for each of the overall goals. We describe goal propagation in two steps. First, we provide an algorithm for propagation of point-value probabilities, and then show how we can propagate probability distributions using Monte Carlo simulation.

1) Propagation of Point-Value Probabilities.: The basis for propagating point-value probabilities in our approach is the algorithm proposed in [7]. Similar algorithms exist for probability propagation in fault trees, e.g., see [5]. We characterize propagation through the rules shown in Figure 4. In the figure, $P(G_i)$ denotes a (point-value) probability of satisfaction for a goal G_i and $P(O_i)$ denotes a (point-value) probability of occurrence for an obstacle O_i. The α and β values in rules (b) and (d) are described earlier in Section III-B. Rules (a)–(d) concern goal–goal propagation. These apply also for obstacle–obstacle propagation; hence, in Figure 4, we do not repeat the rules for obstacles. Rule (e) deals with propagation from a root obstacle to a goal; a dual rule (not shown) is applied for propagating from a root mitigating goal to an obstacle.

As we also stated in Section III-B, for goal–goal propagation, we do not need to elicit β, thus β is set to zero and rule (d) reduces to (c) for goals. In contrast, for obstacle-obstacle propagation, β is important, whereas α is not, hence reducing rule (b) to (a) for obstacles. Lastly, we note that for rules (c) and (d) to apply, all $G_1, \ldots, G_n$ must be simultaneously realized by the system under assessment as alternative ways to satisfy G. In other words, if OR decomposition is used for exploring different alternatives and choosing one (or a subset) of them for realization, then all the unrealized alternatives must be removed before rules (c) and (d) can be applied.

An assumption underlying goal propagation is that the subgoals that a parent goal are elaborated into are independent from one another. This assumption is common [7] and consistent with best practice in argumentation, which is to decouple arguments that are not explicitly related through decomposition. As for obstacles, some may represent common-cause failures [5], e.g., loss of electrical power, flooding, ventilation, and human errors. In such cases, an obstacle can obstruct multiple goals. There are approaches for expressing and quantitatively reasoning about common-cause failures, see [15] for an overview. However, these approaches do not apply at the level of abstraction of goal models, as they require detailed information about the sequence and timing of the failures. In our work, we follow the standard approach in fault-tree analysis [5] and include multiple copies of common-cause obstacles in the goal model. In other words, for a common-cause obstacle O, we include a separate copy of O at every location where O is causing an obstruction. In the random sampling stage of the Monte Carlo simulation (described later in this section), we make sure that in each iteration, only a single random value is drawn for each common-cause obstacle O, and not different values for the different instances of O in the goal model. Given the rules in Figure 4, (point-value) propagation for the entire goal model can be performed using the simple algorithm shown in Figure 5.

2) Propagation of Probability Distributions.: To compute a probability density curve for the satisfaction of an overall goal (or the occurrence of an overall obstacle), we use Monte Carlo simulation [18]. The simulation algorithm is shown in Figure 6(a). To run the algorithm, we need to specify the number of iterations (R). Each iteration begins with the generation of random input variables ($\bar{x}_1, \ldots, \bar{x}_n$) according to the probability distribution for each variable. In our case, the probability distributions for the variables are the triangular distributions elicited from the experts. In the next step, we run the point-value propagation algorithm in Figure 5 and record the resulting value $\bar{y}_i$. After running the algorithm for R rounds, we construct an (approximate) probability density curve for $\bar{y}$ by computing the frequency of the observed values falling into the different value ranges between the min. and max. observed values for $\bar{y}$.

Figure 6(b) shows the probability density curve for goal GL3, with $R = 10,000$. The curve is based on the actual distributions elicited from the experts for GL6, GL7, GL9, GL10, GL11; and α for the decompositions of GL3, GL5, and GL8. We note that, while this could be different in other

Table I
BIAS MONITORING AND MITIGATION GUIDELINES.

Name	Description	Signals	Mitigation strategy
Groupthink	Experts tend to minimize conflicts and reach consensus without critically evaluating others' ideas.	No one voices a difference of opinion; experts appear to defer to other members of the group.	Warn members about intrusion of groupthink. If there is a group leader, solicit their response last or in private. Use anchoring, e.g. have experts write their judgement first.
Wishful thinking	Experts' hopes influences their judgment.	Experts were previously judged to gain something from their answers; experts appear to answer quickly and with little thought.	Have the experts explain their answers in more detail.
Inconsistency	Experts are inconsistent in their solving of problems.	Response mode is applied more easily through time. Extremes of the ratings are being applied as the interviewees get more fatigued.	Avoid fatigue and have the experts review the questions, definitions, assumptions, and response mode.
Availability	Experts retrieve events with different ease from long-term memory.	Experts do not mention more than one or two considerations prior to answering.	Stimulate the expert's memory associations; ask experts to refrain from being critical to generate the widest possible pool of ideas.
Anchoring	Experts rely too heavily, or "anchor," on one trait or piece of information when making decisions.	Experts receive additional information from other experts or sources during elicitation but never waiver from their first impressions.	Ask for extreme judgements before obtaining likely ones; ask experts to describe how other experts might disagree with their responses; ask the experts to temporarily forget recent events.
Overconfidence	Experts underestimate their uncertainty.	Too little uncertainty or variation is expressed while providing answers.	Disaggregate the questions and elicit quantities for finer-grained questions.

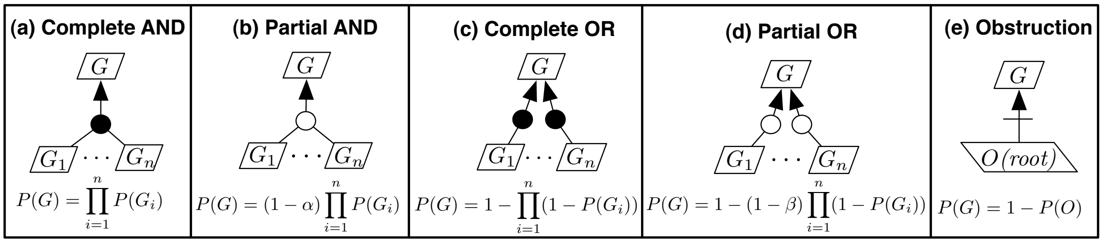
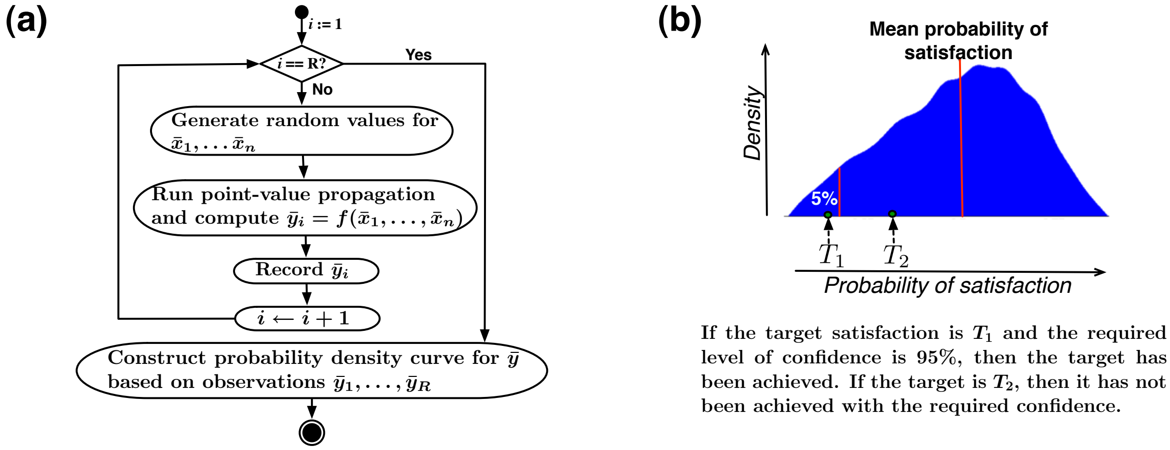


Figure 4. Rules for point-value probability propagation.

```
▷ Mark all leaf goals and obstacles as “done”;
▷ while (tree root is not “done”):
    ▷ for every goal or obstacle $X$ all of whose children are “done”:
        ▷ Apply appropriate propagation rule from Figure 4;
        ▷ Mark $X$ as “done”;
    ▷ end for;
▷ end while;
▷ output value for tree root;
```

Figure 5. Point-value propagation algorithm.


(a)
$i := 1$
$i == R$?
Yes
No
Generate random values for $\bar{x}_1, \ldots \bar{x}_n$
Run point-value propagation and compute $\bar{y}_i = f(\bar{x}_1, \ldots, \bar{x}_n)$
Record $\bar{y}_i$
$i \leftarrow i + 1$
Construct probability density curve for $\bar{y}$ based on observations $\bar{y}_1, \ldots, \bar{y}_R$
(b)
Mean probability of satisfaction
Density
5%
T_1
T_2
Probability of satisfaction

If the target satisfaction is T_1 and the required level of confidence is 95%, then the target has been achieved. If the target is T_2, then it has not been achieved with the required confidence.

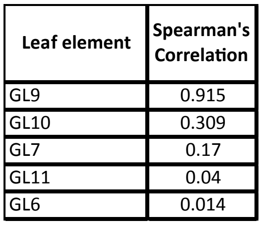
Figure 6. (a) Monte Carlo algorithm (b) Probability density curve for GL3 in Figure 2 situations, the experts in our study provided point-value probabilities for α in all three cases. For privacy, we do not report the elicited quantities, and further, in Figure 6(b) only report the curve shape but not the actual numbers.

In addition to showing the mean probability of satisfaction, the curve provides the level of confidence for the satisfaction. For example, if the targeted probability of satisfaction for GL3 is T_1, the curve tells us, based on the existing evidence, the technology fulfills the target within a 95% confidence interval. In contrast, the technology does not fulfill T_2 within a 95% confidence interval. To reduce the uncertainty associated with the satisfaction of a goal, it is important to be able to identify the main factors that contribute to the uncertainty. This is achieved through sensitivity analysis, discussed next.

D. Sensitivity analysis results.

Sensitivity analysis is typically conducted using one of the following two measures: (1) Pearson's correlation, denoted r, or (2) Spearman's rank correlation coefficient, denoted ρ [4]. These measures, which are computed for each input quantity, indicate how sensitive the output (in our case, the satisfaction of an overall goal) is to that particular input distribution. For Pearson's correlation, quantities should be normally distributed and on interval or ratio scale. Instead, Spearman's rank is non-parametric and only requires ordinal scale. This makes the Spearman's rank more suitable in our context. The higher the correlation between an input and the output, the more significant the influence of the input is on the uncertainty in the output. For example, for GL3 in Figure 2, it is possible to sort the leaf goals based on their impact as shown in Table II.

Table II
SENSITIVITY RESULTS



Leaf element	Spearman's Correlation
GL9	0.915
GL10	0.309
GL7	0.17
GL11	0.04
GL6	0.014

IV. TOOL SUPPORT

We have developed a tool named Modus to support our approach; visit **http://modelme.simula.no/Modus/** for a detailed video demonstration. Briefly, Modus enables users to (1) construct goal models using the KAOS notation and check the models' structural consistency, (2) link and navigate heterogeneous evidence artifacts, (3) perform the expert elicitation steps of our approach and record the elicited probabilities, and (4) export the elicited probabilities and the goal propagation rules as a spreadsheet that can be used for push-button Monte-Carlo simulation and sensitivity analysis in the @Risk risk management tool (http://www.palisade.com/risk/). Modus has been implemented as a plugin for the Enterprise Architect modeling platform (http://www.sparxsystems.com.au), and used with success in our pilot study (Section V). Modus is now being applied in several other ongoing industrial case studies.

V. INDUSTRIAL PILOT STUDY

We have conducted a pilot study in an industrial setting to investigate the feasibility and usefulness of our approach. Our tool (Section IV) was used throughout the study and incrementally refined based on feedback from the experts. The context of the study was the technology qualification program for fiber ropes in the offshore industry. Our study focused specifically on fiber ropes in mooring systems, used for securing and holding floating structures in a fixed position. We used parts of the pilot study for illustrating the various aspects of our approach in earlier sections.

The study was structured in the form of workshops. Three full-day workshops were held for goal-model construction and expert probability elicitation. Five domain experts with background in fiber rope qualification and four researchers participated in the workshops. The researchers acted as facilitators during goal decomposition and expert elicitation activities. In the first step, a high-level goal model consisting of over 80 goals and obstacles was built and validated to express the various safety considerations for fiber ropes. In the next step, the branch dealing with the time-dependent behaviour of fiber ropes (see Section III-A) was further elaborated and linked to the supporting evidence.

Expert elicitation focused on the leaf goals and decomposition nodes of the goal-model fragment indicated with a dashed line in Figure 2. For organizing the expert elicitation sessions, three options were considered: individual interviews, interactive groups, and Delphi meetings. The pros and cons of each option were highlighted to the experts [13]. The experts agreed that an interactive group was the most suitable choice for a first application of a new approach, so that the experts had an opportunity to discuss their views and minimize ambiguity about the quantities under elicitation. After the elicitation activities were concluded, Monte Carlo simulation and sensitivity analysis were performed for goal G3 in Figure 2. Results were earlier discussed in Sections III-C2 and III-D. These results were subsequently presented to the experts. They found the results to be intuitive and in line with their expectations based on the evaluations done using current practices. Two key advantages were noted by the experts: (1) the construction of an explicit goal model and following a fully documented elicitation process improved the transparency of the qualification program; and (2) quantitative analysis is a valuable aid for identifying where expert judgment introduces the most uncertainty and for taking steps to reduce the dependence on expert judgment by provision of more thorough evidence. Initiatives are already underway to develop more detailed evidence collection guidelines where uncertainty was found to be high.

Our case study above focused on mechanical and hardware aspects, and software was not considered during assessment. This limitation was dictated by the need to keep the scope of the study small (as this was the first study of its kind), and the unavailability of the relevant software experts at the time.

While case studies involving software would be essential for a more thorough evaluation, we believe that the core principles of our framework are applicable to software as well. Indeed, the probabilistic reasoning model that we use is closely aligned with the notion of Safety Integrity Levels (SILs) in major software safety standards such as IEC 61508 [9]. This standard specifies four SILs (numbered 1–4) for safety functions, with SIL1 being the lowest and SIL4 – the highest. Each SIL is defined as a range for the average probability of failure on demand for low-demand modes of operation, or the probability of a (dangerous) failure per hour for high-demand or continuous modes of operation. Such probabilistic reasoning requires a careful collection of evidence that would support the reasoning. Evidence collection is an area where software lacks maturity when compared to hardware and mechanical components. However, this is a well-recognized problem and the subject of investigation by several researchers, including some of the authors [23]. In a technical report [12], we demonstrate the application of our framework to a benchmark software-dependent system, but without performing the expert elicitation steps and instead using fictitious probabilities.

VI. Related Work

Our work is inspired by and builds on the concept of goal-based assurance cases, and more specifically safety cases. A safety case is defined as a structured set of evidence-supported arguments to demonstrate that a system is acceptably safe for a given application in a given context [10]. The most adopted framework for safety-case construction is the Goal Structuring Notation (GSN) [10]. GSN enables analysts to define and decompose goals in a similar manner to KAOS [22] – the goal language we use. Our motivations for choosing KAOS were described earlier in Section III-A.

Further, despite being founded on the same principles, there is a subtle but important conceptual distinction between the notion of "goal" in GSN and that in KAOS. Specifically, GSN is concerned with "argumentation" goals, whereas KAOS is concerned with "system" goals (and obstacles). Therefore, in GSN, there can be no OR decomposition in the (final) argumentation goals. Conversely, in the context of our work, OR is necessary for refinement of obstacles and also to denote the situation where several alternative measures (goals) are realized by the system to mitigate a particular risk. With this distinction recognized, our approach can be adapted to work with GSN under modified decomposition semantics.

Recently, a quantitative argumentation framework, Trust-IT [3], has been proposed to measure how justified the claims about the dependability (mainly, safety) of a system are. We share with Trust-IT the motivation for quantitative assessment, but use a different mechanism for quantification. Specifically, the basis of quantification in Trust-IT is Dempster-Shafer theory of beliefs [20]; whereas we use probability theory. In addition to being in line with TQ current practices, the use of probability theory has two main advantages: (1) the existence of proven guidelines for expert elicitation of probabilities [13], [1]; (2) the flexibility offered by probability theory to conduct advanced analyses such as sensitivity (Section III-D).

Goal propagation has been studied in the Requirements Engineering literature for a long time [7], but the focus has been on propagation of point-values. A notable exception [8] concurrent to our work uses a combination of simulation and search-based techniques for analyzing tradeoffs in quantitative goal models. Our work applies the same mathematical ideas for simulation, but is targeted at safety and reliability quantification as opposed to tradeoff analysis. In addition, our work includes tailored expert elicitation guidelines and sensitivity analysis facilities which are not within the scope of [8].

VII. Conclusion

In this paper, we presented a tool-supported approach for qualification of new technology. The main novelty of our work lies in seamlessly combining goal modeling, expert elicitation, and probabilistic simulation for quantitatively assessing the satisfaction of a technology's safety and reliability goals. Our software tool unifies the various aspects of our approach into a coherent implementation. We applied our approach to a pilot study for an offshore-industry application. Results indicate that our approach offers benefits by enhancing the clarity of the arguments through a graphical representation and providing a natural way to reason about goal satisfaction.

Our approach is aimed at providing a solution for probabilistic assessment, which is a regulatory requirement in many contexts such as technology qualification. To be able to trust its quantitative outcomes, it is critical that our approach be complemented and applied in tandem with *qualitative* safety and reliability measures, both to ensure the precision and adequacy of the goal models built, and to properly reflect on the qualitative insights that experts will have inevitably brought to mind in the probability elicitation process.

In the future, we plan to use the observations and lessons learned from our ongoing case studies to further refine and tailor our approach to the needs of technology qualification. The main focus of our work in the near future will be on:

(1) supporting the modeling and the aggregation of different dimensions of decomposition, particularly, behavioural, structural, and process decompositions. For the integration of behavioural and structural dimensions, we are considering existing guidelines in the AADL's Error Model Annex [19] and the AltaRica language (http://altarica.labri.fr/). (2) Supporting quantitative cost and performance comparisons between alternative choices in the technology design phases. Existing literature on trade-off analysis for security [2] and requirements engineering [8] is a promising starting point in this direction.

References

[1] A. O'Hagan et al. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, 2006.

[2] S. Butler. Security attribute evaluation method: a cost-benefit approach. In *ICSE'02*, pages 232–240, 2002.

[3] L. Cyra and J. Górski. Expert assessment of arguments: A method and its experimental evaluation. In *SAFECOMP'08*, pages 291–304, 2008.

[4] J. Devore and N. Farnum. *Applied Statistics for Engineers and Scientists*. Duxbury, 2nd edition, 2004.

[5] C. Ericson II. *Hazard Analysis Techniques for System Safety*. Wiley, 2005.

[6] P. Garvey, editor. *Probability Methods for Cost Uncertainty Analysis*. Marcel Dekker, 2000.

[7] P. Giorgini, J. Mylopoulos, E. Nicchiarelli, and R. Sebastiani. Formal reasoning techniques for goal models. *J. Data Semantics*, 1:1–20, 2003.

[8] W. Heaven and E. Letier. Simulating and optimising design decisions in quantitative goal models. In *RE'11*, 2011. To appear.

[9] IEC 61508: Functional safety of electrical/electronic/programmable electronic safety-related systems, 2005. Intl. Electrotechnical Commission.

[10] T. Kelly and R. Weaver. The goal structuring notation - a safety argument notation. In *Dependable Systems and Networks 2004 Workshop on Assurance Cases*, 2004.

[11] B. Littlewood and D. Wright. The use of multilegged arguments to increase confidence in safety claims for software-based systems. *IEEE TSE*, 33(5):347–365, 2007.

[12] M. Sabetzadeh et al. MODUS: A goal-based approach for quantitative assessment of technical systems. Technical report, Simula–DNV, 2010. http://modelme.simula.no/assets/modus.pdf.

[13] M. Meyer and J. Booker. *Eliciting and analyzing expert judgment: a practical guide*. SIAM, 2001.

[14] M. Morgan and M. Henrion, editors. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge Press, 1992.

[15] A. Mosleh. Interaction between model and data in common cause failure analysis. Technical Report B9-13, U. Maryland, 1989.

[16] Position mooring. DNV-OS-E301, DNV, 2010.

[17] Qualification procedures for new technology. DNV-RP-A203, DNV, 2001.

[18] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2005.

[19] SAE AADL Annex Volume 1: Annex E: Error Model Annex, 2006.

[20] G. Shafer. *A Mathematical Theory of Evidence*. Princeton Press, 1976.

[21] Technology qualification management. DNV-OSS-401, DNV, 2010.

[22] A. van Lamsweerde. *Requirements Engineering: From System Goals to UML Models to Software Specifications*. Wiley, 2009.

[23] R. P. Walawege, M. Sabetzadeh, L. Briand, and T. Coq. Characterizing the chain of evidence for software safety cases: A conceptual model based on the iec 61508 standard. In *ICST'10*, pages 335–344, 2010.

[24] E. Yu. Towards modeling and reasoning support for early-phase requirements engineering. In *RE'97*, pages 226–235, 1997.